AD-A104 511 BEDFORD RESEARCH ASSOCIATES MA F/G 12/1
STATISTICAL GOODNESS-OF-FIT TECHNIQUES APPLICABLE TO SCINTILLAT--ETC(U)
JUN 80 P TSIPOURAS, R D'AGOSTINO F19628-79-C-0163
UNCLASSIFIED SCIENTIFIC-1 AFGL-TR-80-0345 NL

1 OF 1
AD A
104511

END
DATE
FILMED
10-81
DTIC

LEVEL II (12) BS

AFGL-TR-80-0345

STATISTICAL GOODNESS-OF-FIT TECHNIQUES
APPLICABLE TO SCINTILLATION DATA

P. Tsipouras
R. D'Agostino

Bedford Research Associates
2 DeAngelo Drive
Bedford, Massachusetts 01731

Scientific Report No. 1

30 June 1980

Approved for public release; distribution unlimited

AIR FORCE GEOPHYSICS LABORATORY
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
HANSCOM AFB, MASSACHUSETTS 01731

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
AFGL-TR-80-0345

2. GOVT ACCESSION NO.
AD-A104511

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Statistical Goodness-of-fit Techniques Applicable to Scintillation Data

5. TYPE OF REPORT & PERIOD COVERED
Scientific [redacted] No. 1

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

P. Tsipouras*
R. D'Agostino

8. CONTRACT OR GRANT NUMBER(s)
F19628-79-C-0163

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Bedford Research Associates
2 DeAngelo Drive
Bedford, MA. 01730

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
62101F
9993XXXX

11. CONTROLLING OFFICE NAME AND ADDRESS
Air Force Geophysics Laboratories
Hanscom AFB, Massachusetts 01731
Monitor/Paul Tsipouras/SUWA

12. REPORT DATE
30 Jun [redacted]980

13. NUMBER OF PAGES
20

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

* Air Force Geophysics Laboratories
Hanscom AFB, MA. 01731

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Nakagami-m, Goodness-of-Fit tests, probability distributions.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Nakagami-m distribution is often suggested as the probability distribution appropriate for scintillation data. This report considers the statistical questions related to judging the goodness-of-fit of data to the Nakagami-m distribution. Both graphical and numerical techniques are discussed.

1. Introduction

Satellite communication links at UHF can be subject to the effects of ionospheric scintillations. These scintillations cause both enhancements and fading about the median level as the radio signal transmits the disturbed ionospheric region. When scintillations occur which exceed the fade margin, performance of the communications link will be degraded. Of major importance is the estimation of the occurrences of these scintillations that result in degradation of the communications link. One approach to this estimation problem consists of determining the probability distribution of scintillations and then using the properties and parameters of this distribution to obtain the desired estimates related to fading. The Nakagami-m distribution (Nakagami, 1960) has been shown to be a useful distribution for describing the effects of scintillations (Whitney, Aarons, Allen and Seeman, 1972). This paper discusses the statistical procedures that are applicable in attempting to judge the goodness-of-fit, or appropriateness of the Nakagami-m distribution to a data set. That is, it discusses procedures for determining if a given data set can be considered a sample of data generated from a Nakagami-m distribution.

While the underlying problem which this paper addresses did arise from an investigation of scintillation data, the statistical techniques are not specific to this problem. The techniques are applicable to any data set arising as either independent observations or as a stationary time series which might be from a Nakagami-m distribution.

2. Mathematical Properties of the Nakagami-m Distribution

The probability density for the Nakagami-m distribution is normally given as an amplitude or power probability density function

$$f_S(s) = \frac{m^m}{\Gamma(m)\Omega^m} s^{m-1} \exp\left(-\frac{ms}{\Omega}\right) \quad (2.1)$$

where

s = signal power (watts),

Ω = average power,

$1/2 \leq m \leq \infty$

and

$\Gamma(m)$ = gamma function of m.

For modelling scintillation data m=1 in (2.1) is referred to as Rayleigh fading. In such a situation the probability density is

$$f_S(s) = \frac{1}{\Omega} \exp\left(-\frac{s}{\Omega}\right) \quad (2.2)$$

This distribution (2.2) is usually called the exponential or negative exponential distribution. (2.2) is related to the classical Rayleigh distribution when we consider the transformation of variables

$$R^2 = S \quad (2.3)$$

where R is intensity. The probability density of R is then

$$f_R(r) = \frac{2r}{\Omega} \exp\left(-\frac{r^2}{\Omega}\right) \quad (2.4)$$

which is the classical Rayleigh distribution. Also for scintillation data m 1 refers to fading more severe then Rayleigh fading.

The nth moment of the random variable S of (2.1) about the origin is

$$ES^n = \overline{S^n} = \int_0^\infty s^n f_S(s)\, ds = \left[\left(\frac{m^m}{\Gamma(m)\,\Omega^m}\right) \frac{(n+m)}{\left(\frac{m}{\Omega}\right)^{(n+m)}}\right] \qquad (2.5)$$

For scintillation data an important parameter is the coefficient of variation or, in scintillation jargon, the S_4 index. Here

$$S_4 = \frac{(ES^2 - (ES)^2)^{1/2}}{ES} = \frac{\sigma}{\mu} = \frac{1}{\sqrt{m}} \qquad (2.6)$$

In (2.6) we use μ and σ to represent, respectively, the mean and standard deviation. m is the parameter of the Nakagami-m distribution given in (2.1).

3. Relation to the Gamma Distribution

The Nakagami-m distribution as given in (2.1) is related to the more standard gamma distribution whose density is given by

$$f_S(s) = \frac{s^{\alpha-1}}{\Gamma(\alpha)\lambda^{\beta}} \exp\left(\frac{S}{\lambda}\right) \qquad (3.1)$$

Notice if we set in (3.1)

$$\alpha = m \text{ and } \lambda = \Omega/m \qquad (3.2)$$

the gamma density of (3.1) is equal to the Nakagami-m density of (2.1). This relationship is important for it allows us to use the extensive theory developed for the gamma distribution to solve problems dealing with the Nakagami-m distribution.

4.1 Graphical Analysis

4.1 Ecdf and Kolmogorov-Smirnov Test

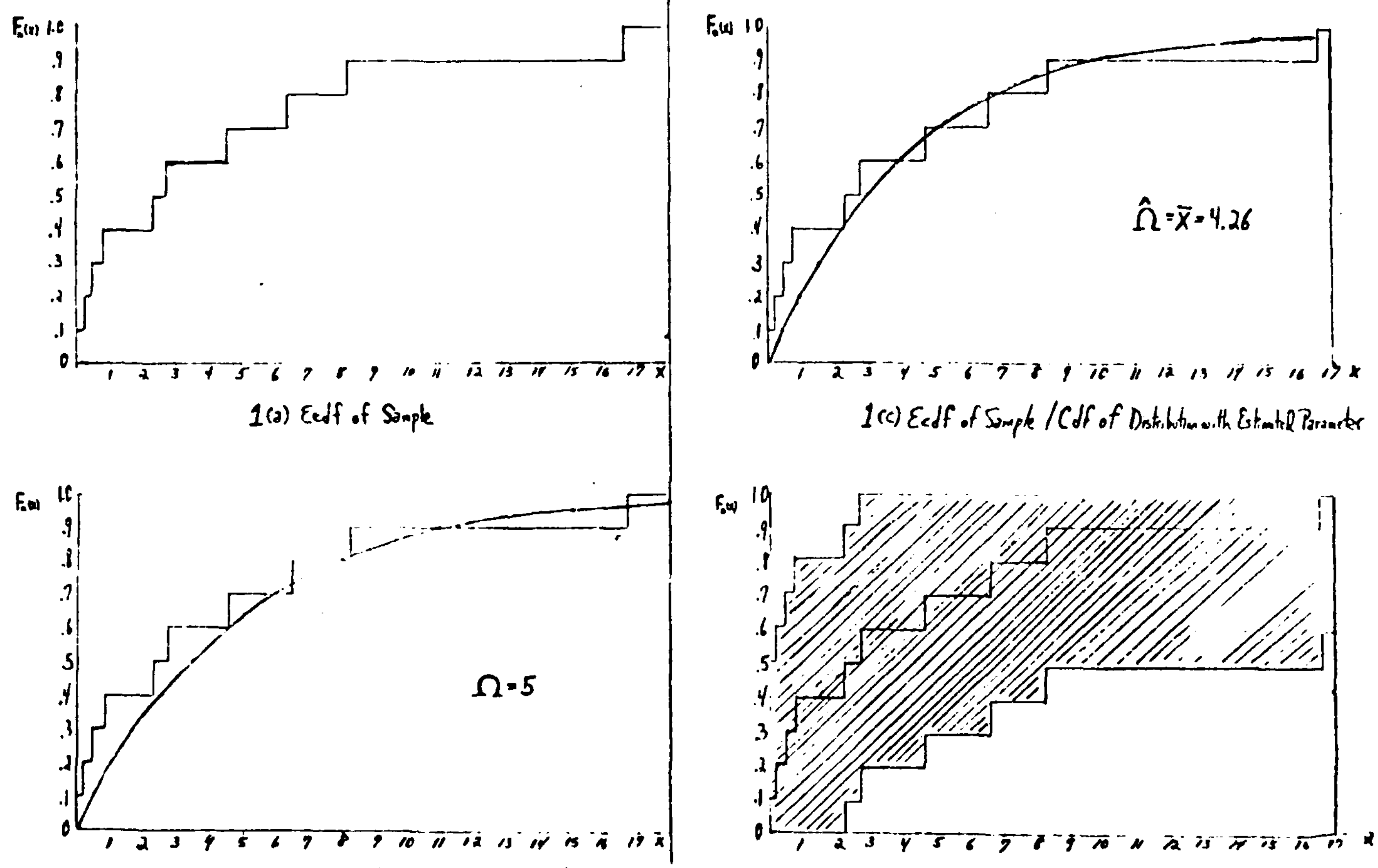
Say $X_1, \ldots, X_n$ represent a sample of size n. Further, say we wish to evaluate whether this sample came from a Nakagami-m distribution. To begin the analysis one should first compute the empirical cumulative distribution function (ecdf) defined for arbitrary x as

$$F_n(x) = \frac{\#(X_i \leq x)}{n} \tag{4.1}$$

Figure 1a contain an ecdf plot of 10 random observations drawn from a negative exponential distribution with mean 5 (i.e., Rayleigh fading, m=1 and $\Omega = 5$ in (2.1)). The 10 observations are: 8.15, 4.69, 2.17, 0.37, 16.69, 0.06, 6.48, 2.63, 0.44, 0.89. The sample arranged in order of magnitude and with the corresponding ecdf values are:

Ordered Observation Number (i)	Ordered Observation	Ecdf $F_n(x) = \frac{i}{n}$
1	0.06	.10
2	0.37	.20
3	0.44	.30
4	0.89	.40
5	2.17	.50
6	2.63	.60
7	4.69	.70
8	6.48	.80
9	8.15	.90
10	16.69	1.00

The next step is to plot the cumulative distribution function (cdf) for the hypothesized distribution on the same graph with the sample ecdf and then judge if the ecdf differs significantly from the hypothesized cdf. For a continuous random variable X with probability density F(x) the cdf is defined as



1(a) Ecdf of Sample

1(c) Ecdf of Sample / Cdf of Distribution with Estimated Parameter

1(b) Ecdf of Sample / Cdf of True Distribution

1(d) Ecdf of Sample with .95 Confidence Band

Figure 1. Graphical Analysis of Exponential Sample

$$F(x) = \int_{-\infty}^{x} f(y)\, dy \tag{4.2}$$

Say for the present example the hypothesized distribution is the exponential given in (2.2). The cdf for this distribution is

$$F(x) = 1 - e^{-x/\Omega} \qquad \text{for } x \geq 0$$

and

$$F(x) = 0 \qquad \text{for } X < 0 \tag{4.3}$$

Recall this distribution represents Rayleigh fading (i.e, m=1 in (2.1)).

Two situations present themselves here. First, the values of all the parameters of the hypothesized distribution are known. For our example the only parameter is Ω. Figure 1b contains, in addition to the ecdf of the observations, the cdf of (4.3) with Ω=5. The second situation is when the values of some of the parameters are not known. Figure 1c contains, in addition to the ecdf of the ten observations, the cdf of (4.3) where Ω is replaced by an estimate, $\hat{\Omega}$, of it which is the sample mean $X = \bar{4}.26$. In general the unknown parameters should be replaced with efficient estimates - e.g., minimum variance estimators or maximum likelihood estimators. However, estimators obtained by the method of moments are also often used for "quick computations". For our example the moment estimator is also the minimum variance and maximum likelihood estimator.

To judge the significance of the difference between the cdf and the ecdf the investigator has two possibilities. The first is simply to judge informally if the difference is too large. For example, the investigator can compute

$$F_n(x) - F(x) \tag{4.4}$$

for a variety of x's and make a judgement concerning their magnitudes. In this situation the investigator is usually asking the question "Are the differences in (4.4) of any practical significance? The second procedure consists of using a formal statistics test of significance - viz., the Kolmogorov-Smirnov test (see Dixon and Massey, 1969, p. 345). This test consists of computing

$$D = \sup_{x} \left| F_n (x) - F_{(x)} \right| \tag{4.5}$$

and rejecting the hypothesized distribution as the true distribution if D of (4.5) exceeds a critical value, say d_α. The value d_α is selected to produce a test of level of significance equal to α- i.e., it is selected so that there is an α chance of $D \geq d_\alpha$ if the hypoth4sized distribution is the true distribution. Alternatively this test consists of adding d_α to all values of $F_n(x)$. The results of such a computation are shown in Figure 1d. (Note in Figure 1d $F_n (x) \pm d_\alpha$ is forced to lie between 0 and 1. Also for this figure $d_\alpha = .41$ for $n = 10$ and $\alpha = .05$). If any of the cdf is outside the band, the hypothesized distribution is rejected as the true distribution at the α level of significance.

This last version of the Kolmogorov-Smirnov test can also be used to produce a confidence interval or region for the underlying distributions cdf. Any cdf lying completely in the band $F_n (x) \pm d_\alpha$ is an acceptable cdf at the 100 $(1-\alpha)$ percent level of confidence. For example, the shaded area in figure 1d consists of the 95 percent confidence region for the 10 random observations given above.

The values d_α of the Kolmogorov-Smirnov test depend upon the desired level of significance (or desired confidence level) and the sample size. One table of d_α values is given in Dixon and Massey (1969). When the sample size of n independent observations exceeds 30 the following values of d_α may be used:

Significance Level	Confidence Level	$d\alpha$
.10	.90	$1.22/\sqrt{n}$
.05	.95	$1.36/\sqrt{n}$
.01	.99	$1.63/\sqrt{n}$

4.2 Special Considerations for Application of Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test as described above is applicable to situations where we have samples consisting of independent observations and all the parameters of the hypothesized distribution are given explicitly. If the observations are not independent, as will happen when we have a time series, then we have two possible procedures. First, the test can be applied to only a subset of the observations. One way of obtaining this subset is to compute the autocorrelation function, find the period or lag that corresponds to a zero autocorrelation (say it is period k) and use every kth observation in the Kiomogorov-Smirnov test. Alternatively, the ecdf of (4.1) can be computed using all the observations, but the values of d_c should be multiplied by $\sqrt{k}$. This will effectively reduce the sample to n/k independent or uncorrelated observations without the loss of any information obtainable from the full set of n available observations.

If any of the parameters are not known then they must first be estimated before the Kolmogorov-Smirnov test can be applied. For the Nakagami-m distribution of (2.1) the two parameters that need to be estimated are

$$\Omega \text{ and } m$$

These can be estimated by the method of maximum likelihood or, if the sample is large, there should be little loss in efficiency if the parameters are estimated by the method of moments. The moment estimates are:

$$\hat{\Omega} = \bar{x} \tag{4.6}$$

and

$$\hat{m} = \frac{\bar{x}^2}{s^2} \tag{4.7}$$

In (4.5) and (4.6) $\bar{x}$ and s^2 are the sample mean and variances, respectively, where

$$\bar{x} = \frac{\Sigma x}{n} \qquad s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

When the parameters are estimated and then the Kolmogorov-Smirnov test is applied, the resulting test is conservative. That is, the true level of significance is smaller than the nominal or stated level.

4.3 Probability Plotting

If the above procedure leads to rejection of the hypothesis that the Nakagami-m distribution "fits" the data, the next item in the analysis is to determine where and how the model deviates from the data. Certain deviations may not be considered to be of practical significance. For example, it may not be a serious lack of fit if the data deviates from the model only for the tail observation (say, less than 2nd percentile or greater than 98 percentile). However, other deviations may be considered very serious and would render the Nakagami-m model useless. It is important that the investigator knows the statistically significant deviations and knows if they are of practical importance.

One useful way of determining where the model deviates from the data is to employ probability plotting. Probability plotting for the present problem is the plotting of the ordered observations from a sample versus the inverse of the cdf of the Nakagami-m distribution of (2.1). Usually Ω is set equal to unity in the plotting. Specifically, say $x_{(1)} \leq \cdots \leq x_{(n)}$ represents the sample ordered from the smallest to largest. Next, say the Nakagami-m distribution with Ω=1 has cdf (G(z) equal to

$$G\ (z) = \int_{o}^{z} \frac{m^{m}}{-(m)}\ y^{n-1} \exp\ (-my)\ dy \qquad (4.8)$$

Further say we redefine the ecdf to be

$$F(x_{(k)}) - \frac{i-1/2}{n} \qquad (4.9)$$

The probability plotting is a plot on linear-by-linear paper of

$$x_{(i)} \text{ on } G^{-1}(F\ (x_{(i)}) \qquad (4.10)$$

If the Nakagami-m distribution is the "true" distribution the plot given by (4.10) is, within sampling fluctuations, a straight line through the origin with slope equal to Ω. Deviations from a straight line indicate where the deviation from the model exists.

Notice in order to implement the above m must be known.

One possibility for determining m if it is not known a priori is to use the estimator $\hat{m}$ given by (4.7). Also note that the inverse of the cdf defined in (4.8) cannot be given in closed form. However, Wilk, Gnanadesikan and Huyett (1962) supply tables and an outline for a computer program which can be used to obtain these inverses.

5. Chi-Square Goodness-of-Fit Test

We recommended the Graphical Analysis coupled with the Kolmogorov Smirnov test described above as the preferred technique for determining the appropriateness of the Nakagami-m distribution to a set of data. However, there may be situations where the Kolmogorov-Smirnov test may not be applicable. For example, the researcher may know a priori that the extreme tails of the data (below 2nd percentile and above 98th percentile) will not be well approximated by the Nakagami-m distribution. This could be due to accuracy limitations of the measurement instrument. In such a situation the investigator may want to censor (i.e., remove) the tails of the data and not enter these into a formal statistical inference test. The ecdf defined in (4.1), the plot of the ecdf (such as in figure 1a), and the probability plotting as described in section 4.3 are still valid and useful. However, the Kolmogorov-Smirnov test is not valid on censored data. The chi-square goodness-of-fit test is appropriate in this situation as a statistical inference test. We suggest the test should be performed as follows:

(1) Decide upon appropriate values of m and Ω. These may be known as apriori or estimated from the data. The appropriate method for obtaining these is by use of the method of maximum likelihood on the censored data. However, if the sample is large the method of moment estimates (see (4.6)) and (4.7)) should be sufficient.

(2) Using the values of m and Ω obtained from (1) find the values $S_0, \ldots S_1, \ldots S_{20}$ which divide the distribution (2.1) into 20 equal probability sections. Note $S_0=0$, S_1 is determined such that

$$.05 = \int_0^{S_1} f_S\,(s)\,ds$$

S_2 is determined such that

$$.10 = \int_0^{S_2} f_S(s)\, ds_1 \ldots$$

and $S_{20} \to \infty$. This step produces 20 intervals or categories each with probability .05.

(3) Compute the expected values for each of the 20 categories. These will all equal n(.05).

(4) Classify each observation into one of the twenty categories. The frequencies in these 20 categories can be represented by $f_1, \ldots, f_{20}$ where $n = \Sigma f_i$.

(5) Compute the chi square statistic

$$X^2 = \Sigma(f_i - .05n)^2/(.05n) \qquad (5.1)$$

(6) Compare the X^2 value of (5.1) with the appropriate critical chi square value obtained from the chi square distribution with 19 degrees of freedom if m and Ω were estimated, or obtained from the chi square distribution with 17 degrees of freedom if both m and Ω were estimated.

If the data contains dependent observations (as in a time series) then the X^2 value of (5.1) should be divided by k, i.e.,

$$X^2_{new} = X^2/k$$

where k is the period or lag corresponding to a zero correlation in the orogInal data (see section 4.2). The X^2_{new} value is now compared to the critical chi square values obtained from the chi square tables.

6. Test for Changing m Values

One problem which the authors have had to address when attempting to evaluate the appropriateness of the Nakagami-m to time series is the problem of changing m values. That is, while the Nakagami-m distribution may be an appropriate model for the data, the actual value of m is not constant over the entire data set. In some of these situations, the number and locations of the segments that have different m values may be known. In this section, we present a large sample test which can be used to test the equality of the Nakagami-m values for **t** segments of data.

6.1 Mathematical Statement of the Problem

Say we have t sets of independent observations each from a Nakagami-m distribution. The m values for the segments are

$$m_1, m_2, \ldots, m_t$$

The problem is to test for the equality of these m values. That is, we want to test the hypothesis

$$H: m_1 = m_2 = \ldots = m_t = m. \tag{6.1}$$

Equivalent to this hypothesis is the hypothesis

$$H^1: \ S_{41} = S_{42} = \ldots = S_{4t} = S_4 \tag{6.2}$$

where S_{4i} for $i = 1, \ldots, t$ is the S_4 value defined in (2.6) for the i^{th} segment. (Recall $S_4 = 1/\sqrt{m}$.). The test we present in the following tests directly the hypothesis H^1 of (6.2).

6.2 Notation - Statistical Results

Say we have n_i independent observations from segment i for i=1, ..., t. From each segment we compute the sample mean, sample standard deviation, and sample estimate of S_4. These are, respectively,

$$\bar{X}_i = \Sigma\,(X_{ij})/n_i \tag{6.3}$$

$$S_i = \sqrt{\frac{\Sigma(X_{ij} - \bar{X}_i)^2}{n-1}} \tag{6.4}$$

and

$$\hat{S}_{4i} = \frac{s_i}{\bar{X}_i} \tag{6.5}$$

Here X_{ij} represents the j^{th} observation for the i^{th} sample, j-1, ..., n_i, i=1, ..., t. For large samples

$$E\hat{S}_{4i} = \frac{1}{\sqrt{m_i}} \tag{6.6}$$

where E represents the expected value operator. Further for large samples the standard error of $\hat{S}_{4i}$ for i=1, ..., t is

$$\sigma_{\hat{S}_{4i}} = \left[\frac{1}{n_i} \quad \frac{1}{m_i} \left[\frac{\mu_{4i} - \mu_{2i}^2}{4\mu_{2i}} + \frac{\mu_{2i}}{\mu_i^2} - \frac{\mu_{3i}}{\mu_{2i}\,\mu_i} \right] \right]^{1/2} \tag{6.7}$$

where

$$\mu_i = EX_{ij} \tag{6.8}$$

and

$$\mu_{\ell i} = E(X_{ij} - \mu_i)^{\ell} \text{ for } \ell \geq 0 \tag{6.9}$$

The sample estimate of the standard errof of $\hat{S}_{4i}$ is

$$\sigma_{\hat{S}_{4i}} = \left[\frac{1}{n_i} \quad (\hat{S}_{4i})^2 \left[\frac{\hat{\mu}_{4i} - \hat{\mu}_{2i}}{4\hat{\mu}_{2i}} + \frac{\hat{\mu}_{2i}}{\bar{X}_i^{\ 2}} - \frac{\hat{\mu}_{3i}}{\hat{\mu}_{2i}\bar{X}_i} \right] \right]^{1/2} \tag{6.10}$$

Here

$$\mu_i = \frac{\Sigma(X_{ij} - \bar{X}_i)}{n_i} \quad \text{for } i = 1, \ldots, t.$$

Further for large samples the $\hat{S}_{4i}$ are approximately normally distributed. See Rao (1973, Chapter 6) for proofs of the above assertions.

If the m_i values are all equal (that is, hypothese H of (6.1) and H^1 of (6.2) are correct, then an estimate of the common value m is

$$\hat{S}_4 = \frac{\Sigma(n_i \hat{S}_{4i}/\hat{\sigma}^2_{\hat{S}_{4i}})}{\Sigma(n_i/\hat{\sigma}^2_{\hat{S}_{4i}})} \tag{6.11}$$

6.3 <u>The Test</u>

Given that H^1: $S_{41} = \ldots = S_{4t}$ is correct then

$$\Sigma_i \frac{(\hat{S}_{4u} - \hat{S}_4)^2}{\hat{\sigma}^2_{\hat{S}_{4i}}} \tag{6.12}$$

is approximately distributed as a chi square variable with t-1 degrees of freedom for large samples. Rejection of H^1 at the α level of significance follows if the statistic of (6.12) exceeds the upper α value of the chi-square distribution with t-1 degrees of freedom (Rao, 1973, p. 389).

6.4 Further Comments

In addition to the m values varying from segment to segment the Ω value of the Nakagami-m distribution may also vary. This value is the mean of the distribution so an appropriate test to test the hypothesis

$$H^*_0 : \Omega_1 = \ldots = \Omega_t$$

is the analysis of variance test. Because the mean of the Nakagami-m distribution is proportional to its standard deviation the analysis of variance on the logs of the data may be a more appropriate analysis than an analysis of variance of the original data (Dixon and Massey, 1969, Chapter 16).

As was stated a number of times above, the test for equality of the m values assumes independent observations. If we are dealing with a time series then the sample sizes should be reduced or other adjustments should be made to reflect this (see section 4.2). One possibility is to use every k_i^{th} observation in the i^{th} segment for i-1, ..., t where k_i is the period or lag corresponding to zero autocorrelation for the i^{th} segment. Then only n_i/k_i observations will be used in each segment. Alternatively, all n_1 observations can be used by n_i in formulas (6.7) and (6.10) should be replaced by n_k/k_i.

REFERENCES

Dixon, W. J. and Massey, F. J., Introduction to Statistical Analysis, (3rd Edition), New York, McGraw-Hill Co., 1969

Nakagami, M., The m-distribution - a general formula of intensity distribution of rapid fading, in Statistical Methods in Radio Wave Propagation, W. C. Hoffman, Editor, pp. 3-36, New York, Pergamon, 1960.

Rao, C. R., Linear Statistical Inference and Its Applications, New York, Wiley, 1973.

Wilk, M. B., Gnanadesikan, R. and Huyett, M. J., Probability plots for the gamma distribution, Technometrics, 4, 1-20, 1962

Whitney, H. E. Aarons, J., Allen, R. S., and Seeman, D. R., Estimation of the cumulative amplitude probability distribution function of ionospheric scintillations, Radio Science, 7, 1995-1104, 1972

DATE

ILME